2015 IEEE 15th International Conference on Data Mining Workshops

Home Location Inference from Sparse and Noisy Data: Models and Applications

Tianran Hu, Jiebo Luo, Henry Kautz and Adam Sadilek *
Department of Computer Science
University of Rochester
Rochester, NY 14623
{thu, jluo, kautz, sadilek}@cs.rochester.edu

***Abstract*—Accurate home location is increasingly important for urban computing. Existing methods either rely on continuous (and expensive) GPS data or suffer from poor accuracy. In particular, the sparse and noisy nature of social media data poses serious challenges in pinspointing where people live at scale. We revisit this research topic and infer home location within 100 by 100 meter squares at 70% accuracy for 71% and 76% of active users in New York City and the Bay Area, respectively. We believe this is the first time home location is detected at such a fine granularity using sparse and noisy data. Since people spend a large portion of their time at home, our model enables novel applications that were previously impossible. As a specific example, we focus on modeling people's *health* at scale.**

Home, as one of the most important locations in people's mobility patterns, is the key to understanding many aspects of urban life and environments. With the knowledge of where people actually live, researchers are able to model the distribution of population, study human mobility patterns, infer life styles, and even discover the correlation between home location and other important aspects such as health conditions, disease diffusion and environment changes.

Much of the research in above mentioned areas is based on surveys and census, which are costly and often incur a delay that hamper real-time analysis and response. Fortunately, the wide adoption of geotagged social media provides us a new opportunity to feel the pulse of cities. In this paper, we present a machine learning based approach to detecting home locations at the population level only based on geo-tagged tweets and use the estimated home locations to investigate these crucial aspects of urban life.

The practicability of a home detection method for urban studies depends on two metrics. The first one is granularity, which indicates in what resolution a method can predict one's home; the second one is applicable scope, which measures the ratio of population that a method is applicable to. In some prior work, granularity are also called "resolution" [1]. In this paper, we use these two terms interchangeably. To look deep into city life, an acceptable method should not only precisely determines ones' home, but also cover as many people as possible.

Significant work has been done to find where people live based on a wide variety of data sources, such as GPS data [2]–[4], cellphone recording data [5] and geotagged social network data [6]–[8]. High quality data such as continuous GPS data and diary data were required to reduce the possible range of one's home location. Krumm et al. [4] reported that home can be located with a median error of about 60 meters using GPS traces of vehicles. However the difficulties in collecting GPS data leads to the low applicable scope of these type of methods. The wide adoption of social media can help us overcome the low applicable scope problem, but much of the existing work could only locate home at a low resolution (city level, state level or even time zone level) based on social media data [9].

In this paper, we investigate ways to balance granularity and applicable scope. In most of the previous work based on geo-attached data, home was simply estimated as the place where one visited most frequently (most check-in place) [5]–[8]. We will show that this method does not always work especially when a user visits several places with similar frequencies. In contrast, we extract the features of one's trajectory in terms of temporal, spatial and other aspects from a Twitter user's sparse trace of locations based on the geotagged tweets. A machine learning method is employed to capture the inherent properties of home using these mobility features, and further detect one's precise home location. We evaluate our method on two large Twitter data sets from the Great New York City(NYC) Area and the Bay Area and the results show that our method is capable of locating homes within a 100 by 100 meter square with a 70% accuracy and applicable to 71% and 76% active Twitter users in NYC and the Bay Area, respectively. An active Twitter users is defined as one who sent at least 5 geo-tagged tweets under using the same definition as in [1], [10], [11]. Utilizing the rich text content within tweets, we explore the health conditions of people in different zip code districts. Note that, the zip code district in NYC has an average area of 3.6 km^2, the radiuses of many of them are less than 1 km. Therefore, finer granularity methods are required. As we will show, our results correlate well (r=0.473, p-value=0.006) with the data from the New York City Department of Health and Mental Hygiene (NYC DOHMH).

I. Related Work

A. Home Location-based User Behavior Understanding

Home location is crucial for modeling human mobility patterns. With the knowledge of home locations, we can gain a better insight of the mobility patterns, as well as lifestyle in general. In [5], [6], [12], home location is the key origin to calculate the distance that people travel and estimate the distance between social network users in a pairwise fashion. Researchers found that home location, as a crucial personal location, can be inferred from information user posted online

* Adam Sadilek is now at Google.

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.149

at a certain granularity [4], [7], [8]. Home location was also used to model individuals' living conditions and lifestyles by Sadilek et al. [13].

B. Home Location Detection

1) Using social network data: In [9], Mahmud et al. used Twitter data attached with geo-information, especially tweet content to infer the home locations at city, state and time zone levels. In this work, their accuracies were 58%, 66%, 78% at city level, state level and time zone level, respectively. Pontes et al. [7] developed "single-attribute" models based on different social network features, for example, taking the value of user's "Employment" as their home city in Google+. They inferred the user home city using geotagged data from Foursquare, Google+ and Twitter with 67%, 72% and 82% accuracy, respectively. In [8], Pontes et al. also used geotagged social network data (Foursquare) to infer the home city within 50 kilometers. A content-based method was used by Cheng et al. [14] to detect Twitter users' home cities. They could place 51% of active Twitter users within 100 miles of their actual home locations. Cho et al. used a data set containing the traces of 2 million mobile phone users from a European country to estimate the home locations according to the most check-ins places in [5]. They reported that by manual checking, the most check-ins method can achieve 85% accuracy when they divided the area into 25 by 25km cells. Scellato et al. [6] simply assigned the most check-ins places as users' home locations but did not provide experiments to verify the accuracy of their method. The location of a person's home is estimated by fitting a two-dimensional Gaussian to all his locations between 1 A.M. and 6 A.M. by Sadilek et al. [13]. The mean of this Gaussian is taken as the most likely home location. In summary, most of the home detection methods work based on social media data require geotagged informatoin. Though the accuracies reported in above studies are reasonable, the granularity levels are so coarse that these method are not applicable on district or finer level study.

2) Using GPS and Diary data: GPS and diary data are much more dense and continuous than social network location data, which make home detection more precise and easier. However, such data are more difficult to obtain. Most of the work using GPS data suffered from the small number of users. In [4], a device recorded location coordinates every several seconds when the car was moving on 172 subjects' vehicles. The ground truth of home location was filled by the drivers themselves. They used 4 heuristic algorithms to compute the coordinates of each subject's home, and found that the best one is "last destination of a day". We also include this feature in our extracted mobility features. The median distance error of their best algorithm was 60.7 meters. Hoh et al. [3] clustered the GPS traces of users agglomeratively until the clusters reached an average size of 100 meters. Next they eliminated clusters with no recorded points between 4pm and midnight and clusters falling outside the residential areas by manual checking.

II. Methodology

A. Data Set & Pre-Processing

We collected all the geo-tagged tweets sent from the greater New York City area during July 2012 and also those sent from the Bay Area during the summer of 2013 through a vendor. A typical geo-tagged tweet contains the ID of the poster, the exact coordinates from where the tweet was sent, time stamp, and the text content. Due to the inherent noise in the geotags, we split the areas into 100 by 100 meter squares and treat the centre of each square as the target of home detection. We assign each tweet to its closest square, each time a user tweet appeared in a square we say s/he had a check-in in this square. Therefore, the "resolution" of our square based home detection is around 70 meters ($\sqrt{2} * 100/2 \approx 70$ meters). Similar to the previous work [1], [10], [11], we only focus on those "active users" who have sent at least 5 tweets. Also following these studies, we use user's hourly traces (only take one location for each hour in our sampling duration) instead of taking account of every single check-in. If a user's location was not observed in an hour, the location for the corresponding hour is set to "Null"; on the other hand, if a user appeared in several unique squares in a hour, we take the square with the highest number of check-ins as the location of this user in this hour. Typically, the hourly trace T_u of a user U looks like: $T = [Null, L_i, Null, ..., L_j]$. L_i does not have to be different with L_j. The lengths of the hourly traces of all users are the same, equalling to the number of hours of our sampling period. We provide a snapshot of our data set in Table I.

	NYC	Bay Area
# of tweets	2,636,437	3,633,712
Total # of active users	55,237	53,314
# of tweets annotated by AMT	5,000	5,000
# home locations (GT)	1,063	987

TABLE I. Statistics of our data set.

B. Ground Truth

In this study, we rely on tweet content and human intelligence. For some tweets, a human can easily tell where it was sent from. For example, if a tweet said "The view from my office is awesome!" and included a picture of the view from a window, we can tell it was sent from a user's office. Some tweets are obviously sent from home, for example, "finally home!" or "home sweet home". This is the basis for us to design a mechanism to build the ground truth for home location. We polled some faithful Twitter users what they would like to post when at home. Based on their answers, we selected a set of keywords, each of which is likely to be mentioned in the tweets sent from home. This set contains words like "home", "bath", "sofa", "TV", "sleep" and so on. We ended up with a set of 50 unique words and their variants. Next, we used a simple keyword filter to obtain all the tweets that contain at least one of these keywords. From here we relied on human intelligence through crowdsourcing on Amazon Mechanical Turk (AMT) to find the "home tweets", which were sent from home. We gathered these tweets into questionnaires. Each questionnaire contains 5 tweets, where we simply ask "do you think these tweets are sent from home?" and the options include "Yes", "No" and "Not sure". We then posted these questionnaires to AMT. Each questionnaire was answered by three unique workers. We only retained the tweets strictly for which all the three workers thought were sent from home.

C. Models based on Human Mobility Features

To study the inherent property of home, we extract several features of every unique location of one's hourly trace. In this section, we discuss these features in detail. Some of these temporal and spatial characteristics can be used as baseline methods to detect home location (e.g. check-in rate, PageRank score). We will show that although a single feature can be used to detect home location with a reasonable accuracy, it usually covers a limited amount of people. However, combining them appropriately using a machine learning method brings us significant gain in applicable scope.

1) Check-in Rate: As we mentioned in Section 2, taking the place of most check-ins as home is a popular method. We call this method "Most Check-in". Due to the different tweet volumes of users, we do not use the absolute check-in amount. Although check-in based methods work well on GPS data [4], it's not the case when it is employed on Twitter data. Unlike vehicle GPS devices which keep recording the location every several seconds, people only tweet when they feel like to do so. The place with most check-ins definitely is important to a user, but "important" does not necessarily mean it is the home. We found that, the effectiveness of Most Check-in is closely related to how much higher the rate of check-ins of the most check-in place than the second most check-in place. Figure 1 shows that, the accuracy of Most Check-in decreases along with the check-in margin linearly. The accuracy is 70% only when the the margin is significantly higher (50% or higher). Therefore, besides the check-in rate of a place, we also include the margins of check-in rate between a place and its next higher and lower most check-in places. Also, this is the reason for the poor applicable scope of Most Check-in when high accuracy is required under our granularity setting (100 by 100 meter square). Figure 2 is the cumulative distribution. It shows that only about 40% users have the margins which are 50% or larger. The inset of Figure 2 reports the the distribution of check-in rate margin between most check-in place and the second most check-in place.

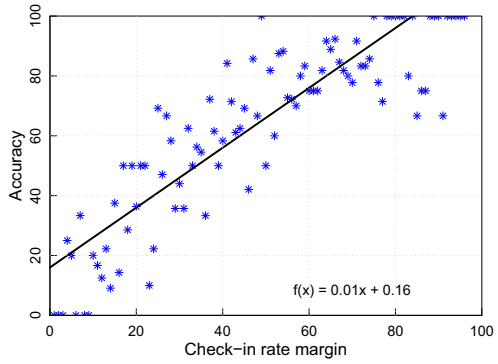

Fig. 1. Accuracy vs. the check-in rate margin between the most check-in place and the second most check-in place.

2) Check-in rate during midnight: Intuitively, the places people appear at midnight are probably their homes. Sadilek et al. [13] took the places with the most check-ins during midnight (00:00-07:00) as people's homes. This method potentially alleviates the biases caused by other frequently visited places during daytime. Therefore, we take the check-in rate during midnight of a place as another feature, separated from the total check-in rate discussed above. However, as we will show later, it is *not* the case among active Twitter users. When an active user checks in at some places during midnight, this place is most probably not the home. This reflects the difference between Twitter data and GPS data. In Twitter data, a user had to be awake and active to report the current location at that moment, while in GPS data, the location recording is automatical even when the users are inactive. We will discuss more on this lately.

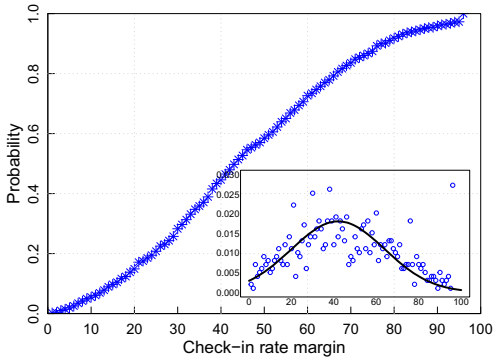

Fig. 2. The cumulative distribution of the check-in rate margin between the most check-in place the the second most check-in place.

3) Last destination of a day: According to the research by Krumm [4] on GPS data, the last destination of a user in a day (no later than 03:00 in the morning) is probably the home. It reveals that people's daily movements end at their homes. We includes this as another mobility feature after minor modification. First, we extract all the final destinations of a user over the entire sampling period. We then sum up the number of days a place had been the final destination of that day. This value, the times of being the last stop of a day, is taken as one of the mobility feature of this place.

4) Last destination with inactive midnight: Since "Last destination" might suffer from the check-ins sent from non-home places especially when the night was spent outside, we also consider a variant feature of last destination. We only consider tweets sent on the days when people were inactive during midnight (00:00-07:00). We exclude the days with active midnight and find the last destination among the remaining days, then count the time of a place of being the last destination.

The three features above introduce extra human behaviour information into the original check-in feature. This helps to reduce the applicable scope limitation of simple check-in rate feature.

5) Spatial Features: As the centre of everyday activities of most users, home is one of the most important starting points and destinations of their movements. We use weighted PageRank [15] and Reversed PageRank scores to model the importance of a place of being an origin point and a destination. To use PageRank related algorithms, we transfer one's trace into a directed graph. Vertexes of the graph are the locations one have visited. A directed edge from location L_i to L_j represents that L_j is visited directly after L_i. To quantify the certainty and importance of transitions between locations, we assign weight to each edge. The weight should be inversely proportional to the length of blank idle between two

locations, and also proportional to the times of a transition appears in one's trace. Formally, let $t(L_i, L_j)$ represent the transition between L_i and L_j and $w_{t(L_i,L_j)}$ represent the weight. The definition of the weight is as follows:

$$w_{t(L_i,L_j)} = \sum_{k_{th}\, t(L_i,L_j)\, in\, T} \frac{1}{\#\, idle\, hours\, in\, k_{th}\, t(L_i, L_j)} \quad (1)$$

After constructing a user's movement graph, we apply the PageRank algorithm to calculate the importance of being a destination for each location in one's trace. Meanwhile, to study the importance of being an origin, we propose a Reversed PageRank score. We reverse each edge's direction in the movement graph, with the weights of edges unchanged. The same reversed calculation is also performed with the weighted PageRank algorithm. Comparing with the earlier features, the PageRank score and Reversed PageRank score describe the spatial characteristics of movements.

6) Temporal Features: According to [4], the probability of being at home varies over time. We extract the check-in rates of a place in different hours. These time related features help us capture the property of home in terms of temporal patterns.

D. Multi-feature prediction

We believe that, as a distinctive place of one's trace, home permeates its influence into all the mobility features discussed above. Indeed, one can use single feature to detect home location. However, a single feature captures only one type of characteristic, and thus will lead to low applicable scope of these methods. We apply a machine learning method to combine all the features. Because of the complementary effect between features, an appropriate combination will significantly increase the method applicable scope without loss of accuracy.

Our goal is to distinguish home from other locations that one has visited. Since we obtain various features for every place of one's trace, the original problem can be transferred to an equivalent classification problem: given locations and corresponding feature values, we want to train a model to predict home among them. The input of the model are transactions identified uniquely by user ID and location ID, followed by features calculated from this user's hourly trace and a label as "home" or "non-home". We use a linear SVM model to exploit how these features are combined. Given the places and their features for a given user, the model outputs a score for each place. If the highest score exceeds a threshold, we take the corresponding place as the user's home. Otherwise, this user cannot be covered by our model. In Table II, we present significantly positive and negative features and their weights. Not surprisingly, check-in rate, PageRank score and Reverse PageRank score related features are more significant than others. Note that all features contribute to the better overall applicable scope.

Positive Features	**Weight**
Check-in ratio	2.03
Margin over second highest check-in	0.19
PageRank Score	0.19
Last destination on inactive midnight	0.12
Reverse PageRank score	0.09
Negative Features	**Weight**
Margin below next higher check-in	-0.30
Margin under next higher PageRank	-0.28
Margin under next higher Reverse PageRank	-0.21
Rank of Reverse PageRank	-0.07
Rank of PageRank	-0.07

TABLE II. SIGNIFICANTLY POSITIVE AND NEGATIVE FEATURES AND THEIR WEIGHTS.

III. HOME LOCATION EVALUATION

To guarantee the practicability of our home detection method, we need to balance granularity and applicable scope. Because of the natural trade-off between granularity and detection accuracy, we fix the granularity as 100 by 100 meter square and explore the relationship between accuracy and applicable scope. The accuracy of each single feature can be adjusted through the threshold, which affects applicable scope as well. In this section, on both NYC and Bay Area Data we compare our method with three other intuitive single-feature based methods: 1) Most Check-in (Due to the statistical insignificance of too few check-ins, we also set a constraint on this method: the absolute check-in number of the most check-in place is at least 3 times.), 2) Highest PageRank Score (Similarly, the threshold is how much higher the highest PageRank Score than the second highest one.) 3) Highest Reversed PageRank Score. Figure 3 and Figure 4 indicates the trend of applicable scope along with accuracy. Applicable scope decreases rapidly as the accuracy goes higher. It shows that, at every accuracy level, our method cover much more users. Especially, when we set the accuracy of each method at 70% (which we think is acceptable for urban computing), our method obtains 71% and 76% applicable scope in NYC and Bay Area, respectively. As to other methods, none of them is able to detect home for more than 50% users when the accuracy is 70% or higher. It proves that an intelligent combination model leads to a significant increase in applicable scope. Most Check-in works better than PageRank Score and Reversed PageRank Score, but the still suffers from low a applicable scope. The higher applicable scope of our method implies the complementary effect between these mobility features. This method dose not depend on any one type of information but combines the information of several aspects. For example, when a user's check-in features do not provide enough cues to predict the home, other types of features may pick up the slack and naturally lead to a higher applicable scope. The balance of applicable scope and accuracy facilitates more extensive and deep urban life studies, which we will describe.

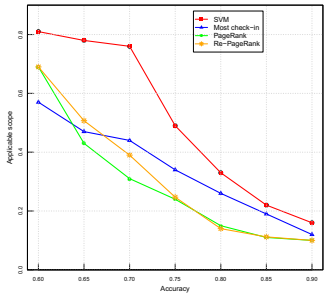

Fig. 3. The applicable scope and accuracies of different methods on NYC data set.

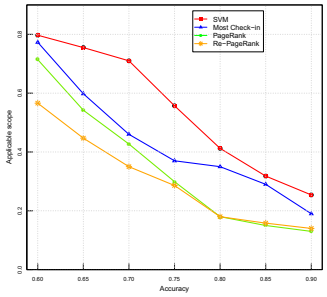

Fig. 4. The applicable scope and accuracies of different methods on the Bay Area data set.

IV. Community Health Condition Estimation

With the precise knowledge of where people live, we are further interested in the relationship between people's health conditions and their home locations. Let's start with how we model the health conditions of tweet users.

A. Inferring Health State

We build upon previous work on classification of short text messages [16]–[18] and learn a support vector machine classier C_s (underscript s indicates sickness to differentiate it from the earlier SVM used for home location) that identifies tweets that indicate their author is ill. C_s is trained by directly optimizing the area under the ROC curve. It is robust even in the presence of strong class imbalance, where for every health-related message there are more than 1,000 irrelevant ones. We use C_s to distinguish between tweets indicating the author is acted by an ailment (we call such tweets "sick tweets"), and all other tweets (called "other" or "normal" tweets). For SVM features, we use all unigram, bigram, and trigram word tokens that appear in the training data. For example, "so sick of" is represented by the feature vector: (so; $sick$; of; $so\ sick$; $sick\ of$; $so\ sick\ of$). As a result, our SVM operates in more than 1.7 million dimensions, where each dimension represents a word or a phrase extracted from the training data. Before tokenization, we convert all text to lower case, strip punctuation and special characters, and remove mentions of user names (the "@"tag) and retweets (analogous to email forwarding). However, we do keep hashtags (such as "#sick"), as those are often relevant to the author's health state, and are particularly useful for disambiguation of short or ill-formed messages. Table III lists examples of significant features found in the process of learning C_s. We use the SVM cascade learning procedure described in [18]. Evaluation of C_s on a held-out set shows 0.98 precision and 0.97 recall with respect to labels agreed upon by human annotators. Ground truth for each tweet was obtained by asking AMT workers to label the tweet as either "sick" or "other" and subsequently extracting the majority vote.

B. Zip Code District Health Condition

We define "sickness score" of a district as the percentage of "sick" people who live in it. Therefore the higher sickness score is, the worse the health condition of this district is.

Positive	Weight	Negative	Weight
sick	0.9579	sick of	-0.4005
headache	0.5249	you	-0.3662
flu	0.5051	lol	-0.3017
fever	0.387	love	-0.175
coughing	0.291	so sick of	-0.08
being sick	0.191	bieber fever	-0.10
better	0.198	smoking	-0.098
being	0.194	i'm sick of	-0.089
stomach	0.170	pressure	-0.083
infection	0.168	i love	-0.071

TABLE III. Examples of positively and negatively weighted significant features of our SVM model C_s.

Area Name	Sickness	Excellent	Good
Upper West Side	0.046	30.7	26.3
Chelsea	0.018	29.3	20
Gramercy	0.029	26.4	19.8
Flatbush	0.100	23.6	30.5
Central Harlem	0.062	23.2	23.9
Lower Manhattan	0.019	23.2	22
Southeast Queens	0.043	22.1	27.6
Astoria	0.042	21	35.1
Crown Heights	0.066	20.8	34.7
Heights/Slope	0.061	20.5	27.1
Inwood	0.049	20.2	38.8
Bushwick	0.070	20	30.5
Southwest Queens	0.084	16.9	33.5
The South Bronx	0.050	13.2	38.6
Fordham	0.083	13	46.1
Pelham	0.085	10.5	38.7
Correlation	NA	**-0.569**	**0.601**

TABLE IV. The sickness score, percentage of people whose health are in Excellent and Good condition. The bottom row shows the correlation between our score and the ratio. The table is sorted by the column "Excellent".

To evaluate our home location method as well as our health inference model, we compare our sickness score with the data from NYC DOHMH. In the data set provided by DOHMH, NYC were divided into 34 areas. The health condition of individuals has 4 levels, "excellent", "very good", "good" and "fairly good". DOHMH provides the percentage of each level of every area. We calculate the correlation between our sickness score and the percentage of "excellent", "very good", "good" level (since the sum of four percentages is one, there is no need to calculate the correlation for all them). Our sickness score is highly negatively correlated with the "excellent" percentage (r=-0.383, p-value=0.030), and positively correlated with "good" percentage of each area (r=0.473, p-value=0.006). This makes sense because our health state inference method is based on the percentage of sick Twitter users, thus it indicates a rough degree of relatively unhealthy people in an area. Intuitively, this degree should be negatively correlated with the percentage of people whose health is in excellent condition. Because there are only four levels in the health survey and the "good" level is the second worst level among the four levels, we consider this level as a "relatively unhealthiness" metric. Therefore, the positive correlation between our sickness score and this level makes sense. Although our method can provide the highest applicable scope among all the method, we still suffer from few sampling problem in some districts. For better data reliability, we excluded all the districts in which there are fewer than 200 residents detected by our method, Table IV shows that the correlations with "excellent" (r=-0.569, p-value=0.017) and that with "good" (r=0.601, p-value=0.010) increase significantly for the districts with sufficient number of residents.

V. Conclusions and Future Work

In this paper, we propose a machine learning based multi-feature method that can precisely locate peoples' home locations. Different from previous works, we do not require continuous GPS trace data but instead utilize noisy and sparse Twitter data. By evaluating on the ground truth obtained by human annotation, our method has achieved 71% and 76% applicable scope on the Twitter data we have collected from New York City and the Bay Area, respectively. To the best of our knowledge, this is the first time that urban life can be studied on such open source data at a fine granularity.

With such a balance, we are able to study human mobility patterns to an extent that was not feasible before.We use a health state model to estimate Twitter users' health conditions. We relate people's health to their home locations and compare our estimated sickness scores with data from NYC DOHMH. Highly correlated results have validated the effectiveness of both our home location estimation method and our health state inference model.

In the future, we will investigate other domains of interest in urban computing given the knowledge of home locations, such as economic activities, resource consumption, urban planing and emergency management. It is also interesting to extend our method for home location to general place recognition based on user movement behaviors.

References

[1] M. Lin, W.-J. Hsu, and Z. Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," in *UbiComp*, 2012, pp. 381–390.

[2] A. Sadilek and J. Krumm, "Far out: Predicting long-term human mobility." in *AAAI*, 2012.

[3] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing security and privacy in traffic-monitoring systems," *Pervasive Computing, IEEE*, vol. 5, no. 4, pp. 38–46, 2006.

[4] J. Krumm, "Inference attacks on location tracks," in *Pervasive Computing*. Springer, 2007, pp. 127–143.

[5] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 2011, pp. 1082–1090.

[6] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks." *ICWSM*, vol. 11, pp. 329–336, 2011.

[7] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida, "Beware of what you share: Inferring home location in social networks," in *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, 2012, pp. 571–578.

[8] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida, "We know where you live: Privacy characterization of foursquare behavior," in *UbiComp*, 2012, pp. 898–905.

[9] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home locations of twitter users." in *ICWSM*, 2012.

[10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[11] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," in *Pervasive Computing*, 2014, pp. 88–94.

[12] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1046–1054.

[13] A. Sadilek and H. Kautz, "Modeling the impact of lifestyle on health at scale," in *WSDM*, 2013, pp. 637–646.

[14] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 759–768.

[15] W. Xing and A. Ghorbani, "Weighted pagerank algorithm," in *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, 2004, pp. 305–314.

[16] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*, 2010, pp. 115–122.

[17] M. J. Paul and M. Dredze, "A model for mining public health topics from twitter," *HEALTH*, vol. 11, pp. 16–6, 2012.

[18] A. Sadilek, H. A. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions." in *ICWSM*, 2012.